DOCUMENT RESUME

ED 189 107 TM 800 258

AUTHOR Moran, Edward; And Others
TITLE Assessing and Improving the Accuracy of Title I Evaluation with Grade-Appropriate Tests. NTS Paper Series.
INSTITUTION NTS Research Corp., Durham, N.C.
PUB DATE Apr 80
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Ability; *Achievement Tests; Compensatory Education; *Difficulty Level; Grade 3; *Item Analysis; Primary Education; Reading Tests; *Scoring; Test Construction; Test Items; *Test Reliability; *Test Selection
IDENTIFIERS *Elementary Secondary Education Act Title I; Rasch Scaled Scores; West Virginia

ABSTRACT

Since grade-appropriate levels of standardized achievement tests must frequently be used in Elementary and Secondary Education Act Title I evaluations, there may be large discrepancies between test difficulty and the achievement level of those being tested, with resultant inaccuracies in the scores. Procedures based on the Rasch scaling model were used to assess the degree to which scores on grade-appropriate reading achievement tests reflected true achievement, and to develop a subset of the item battery that reflected true achievement and that could be linked to national percentile ranks. Comprehensive Tests of Basic Skills reading vocabulary and reading comprehension subtest scores were obtained for 1,000 West Virginia third graders, and submitted to a Rasch scaling analysis using the BICAL computer program. BICAL performed the scaling, assigned item difficulty and student ability values, and generated item-fit statistics for each item and for six student ability groups. Results of the reliability analyses contradicted the original expectations. In conclusion, there was support for testing achievement with a subset of items that was most appropriate for low-achieving students. Similar procedures were recommended for assessing performance of a specific skill. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RECEIVED
ERIC/TM

AERA 1980

American Educational Research Association
Annual Meeting
Boston, Massachusetts
April 6-11, 1980

ED189107

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Baker
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 800258

NTS Paper Series

NTS Research Corporation
2634 Chapel Hill Boulevard
Durham, North Carolina 27707
(919) 493-3451

Assessing and Improving the Accuracy of Title I Evaluation with Grade-Appropriate Tests

Edward Moran, Robert Baker,
Nick Maruhnich

Edward Moran
West Virginia Department of Education

Robert Baker, Nick Maruhnich
NTS Research Corporation

Prepared for the Panel on
"Methods of Program Evaluation"
Annual Meeting, American Educational
Research Association
April 6-11, 1980
Boston, Massachusetts

Introduction

The United States Office of Education (USOE) Title I Evaluation and Reporting System strongly recommends that Title I students be administered tests whose content and difficulty are appropriate for the students' functional level. Underlying this recommendation are two primary assumptions: first, the use of a functional level test will provide a more content-valid assessment of the students' strengths and weaknesses; and, second, the use of a functional level may reduce biases, both random and systematic, in the students' assessment.

Frequently, however, there are substantial economic and social pressures to use test levels in Title I evaluation that are appropriate for the students' grade level. In many cases, these grade-appropriate levels do not reflect the students' functional level. West Virginia is a state where these pressures appear to exist. The statewide testing program requires that every student in grades 3, 6, 9, and 11 be tested in the spring of each year with grade-appropriate levels of the Comprehensive Tests of Basic Skills (CTBS). Practical and political considerations mitigate against double testing, making it extremely unlikely that Title I students in these four grades will be tested with anything other than a grade-appropriate test. In addition, the compulsory use of grade-appropriate tests at these four grades has a "ripple effect" which causes many local education agencies (LEAs) in West Virginia to use grade-appropriate tests at the other grades. To be more specific, West Virginia is using Level 1 of the CTBS for grade 3 and Level 3 for grade 6. Therefore, many LEAs feel forced to use Level 2 in grades 4 and/or 5, even though Level 2 may not reflect their students' functional level. If such pressures to administer tests to Title I students which are not necessarily

at their functional level exist, it would be extremely useful to develop a procedure to help state education agencies (SEAs) and LEAs examine student performance on appropriate test items on the grade-appropriate tests while ignoring the other items and eliminating the generally confounding information they provide.

Most Title I students in West Virginia score below the 40th percentile on the CTBS. For many of the various levels and subtests, students' scores at or below this percentile level represent scores at or below the chance level. That is, a large proportion of the Title I students' scores fall at or below that score which, on the average, would be expected from purely random performance. The following question then arises: For these students, to what extent do their scores reflect measurement error (particularly guessing) as opposed to true achievement? We propose that this question can be tested using item fit statistics based on the degree of congruence between the observed response matrix of low achieving students and the hypothetical response matrix generated by analyzing the observed response matrix using the Rasch (1960) scaling model, a single parameter logistic model. If the fits are good, then we can suggest that even though the group's scores are extremely low and below the chance level, the scores do appear to reflect true achievement for the group since the scores correspond to the model's expectations.[1]

If the fit is poor, it may still be possible to identify and eliminate those items that appear to be the source of incongruence between the observed and hypothetical response matrices. We would anticipate that the bulk of the items that are retained will be those with high p-values (i.e., easy for the

---

[1]An example of this may occur if students work conscientiously, but slowly, on the test items. In this instance, the below-chance scores will reflect achievement and not measurement error; that is, students' scores will reflect their knowledge on the items they attempted.

overall group); the difficulty of these items might be more appropriate for the ability of low achieving students and thus would be less subject to guessing. The score based on this group of "congruent" items should be a much more accurate reflection of true achievement than the subtest score. In turn, Rasch scaling techniques can be applied to equate students' performance on the "congruent" subset to specific score points on the subtest. This means that the national percentile ranks associated with subtest scores can also be associated with subset scores. Even though most students' scores will still equate to percentiles below 40, the students' ranks produced on the congruent subset will be more reliable and less influenced by guessing since their raw scores will be based on the most appropriate subset of items from the subtest. The subset would then be appropriate for Title I evaluation using Model A-2, which links non-normed tests to normed tests, permitting expected posttest performance on the non-normed test to be predicted from pretest performance. As a result, the achievement of Title I students could be more accurately assessed from their performance on selected portions of grade-appropriate levels of the CTBS, since the assessment of their achievement would be based on the subset of items that yielded a high degree of congruence between the Rasch model's predictions and the observed responses of low achieving students.

Related Literature

Traditionally, reliability and content validity have been the most important criteria for sound test construction. Classical test theory states that the longer the test, the more reliable it is. Therefore, it would seem that making shorter tests from longer total test batteries would be counter productive and would decrease reliability. However, practical procedures for

increasing the reliability of tests by using the Rasch model have been suggested and empirically verified using tests for graduate students (Ryan and Hamm, 1976). Using the Rasch model, Ryan and Hamm calibrated items on full-length tests, deleted items which did not fit the model, and developed shorter tests which proved to have greater reliability than the full-length tests. The concept of removing misfitting items is even more relevant to Title I students than graduate students. The graduate students in the study received tests which were appropriate for their functional level. Thus, one would expect most of the items on these graduate student tests to be appropriate or "fit" the model. On the other hand, Title I students usually receive on-level tests which tend to be more difficult than their functional level. Thus, one would expect many of the items to be inappropriate and not "fit" the model. Consequently, one would expect that with Title I students, the potential for removing the measurement error and increasing the reliability of tests would be even greater than that with graduate students, since there would be even more items which are expected not to fit the model and thereby cause measurement error.

Rasch logistic scaling procedures have been used to develop short, individualized tests for public school measurement (Forbes and Ingebo, 1975). In this study a series of tests were developed and matched to different subgroups of students to suit individual student differences of the various subgroups. The tests were reported to have adequate reliability (even considering their shorter lengths) and much greater content validity.

These studies demonstrate that it may be feasible to use Rasch scaling procedures to create a shortened version of the CTBS that should have greater reliability and content validity for Title I students in West Virginia. Furthermore, there is nothing in measurement literature which suggests that there will be any technical problems in accomplishing such a task.

Method and Results

The data source for this study was the scores of 1,000 randomly selected third grade students on the CTBS reading vocabulary and reading comprehension subtests.[2] These students were administered the CTBS as part of the 1979 West Virginia State Testing Program. We submitted the scores from each subtest to a Rasch scaling analysis using the BICAL program (Wright, Mead, and Bell, 1979). The West Virginia State Testing Program uses Level 1, Form S of the CTBS in the third grade; this test has forty items for reading vocabulary, and forty-five items for reading comprehension. The BICAL program not only performed the scaling, assigning difficulty values to items and ability values to students, but also generated a number of item-fit statistics that were used to investigate the degree to which each item was measuring the underlying trait (e.g., achievement in either reading vocabulary or reading comprehension). The basis for these fit statistics is the disparity between the observed response matrix of 0s (incorrect answers) and 1s (correct answers) and the hypothetical response matrix which contains the probabilities that the items will be correctly answered. These disparities can then be used to derive t-statistics and assess whether the magnitude of the disparities is significant.

The BICAL program also generated a table which presented item-fit statistics for each of six groups. These six groups were chosen by the program to be of approximately equal size, with one group containing the respondents with the lowest raw scores, a second group containing the

---

[2]We extend our thanks to John McClure and Charles Duffy of the West Virginia Division of Evaluation and Information Systems for providing us with this data. It should be noted that students' names were deleted from the data tape that was supplied.

respondents with the highest raw scores, and the other groups containing respondents who scored in the intervening ranges. Thus, the degree to which the hypothetical probabilities match the observed responses for each item could be examined on an achievement level by achievement level basis. Clearly, it is preferable to have items which have a reasonably good fit for all six achievement levels, as opposed to items where the observed responses of one or more groups have a poor fit with corresponding hypothetical probabilities. A t-statistic (with five degrees of freedom) was used to test the significance of the differences between the six achievement level groups with respect to the item fit.

Wright, Mead, and Bell suggest several guidelines when examining the item-fit statistics. The overall item fit, which they refer to as "total t-fit," should certainly be considered poor when it exceeds a value of 2.0. The t-statistic which assesses the between group differences with respect to item fit, which is referred to as the "between t-fit," is considered by Wright, Mead, and Bell to have an asymptotic distribution which will be normal with a mean of 0 and a standard deviation of 1.[3] For purposes of establishing a criterion with respect to the "between t-fit" statistic, we selected a value of 3.0, which lies between the .05 t-statistic (with five degrees of freedom) of 2.015 and the .01 t-statistic of 3.365. Items which either had a poor total t-fit (greater than 2.0) or a poor between t-fit (greater than 3.0) were eliminated. Seventeen of the forty reading vocabulary items were retained and twenty-three of the forty-five reading comprehension items were

---

[3]While the same is supposed to hold true of the "total t-fit" statistic, it has been the experience of Wright, Mead, and Bell that the mean and standard deviation can vary substantially from 0 and 1 respectively.

retained. The numbers of the items that were retained, as well as their p-values, between t-fits, and total t-fits, are presented in Table 1.

An examination of the p-values in Table 1 indicates that, as might be expected, the majority of the items retained tend to be "easy" (a p-value of greater than .50). For both reading vocabulary and reading comprehension, however, a number of items had p-values below .50. Thus, through serendipity, a potentially serious problem has been avoided. If only the easiest items had been retained, a test "ceiling" would have been established, making it impossible for a student to get a score which would indicate that his/her achievement level was above the low range. This could seriously impair the capability of the student to show growth with respect to this subset of items, and thus could depress the observed magnitude of the Title I effect.

In order to assess the psychometric characteristics of the reading vocabulary and reading comprehension item subsets, two analyses were performed. First, a reliability analysis (Specht, 1975) was performed. The KR_{20} coefficient was computed for the entire subtest and for the subset for all students and for those students scoring below the 40th percentile. The results for reading vocabulary are summarized in Table 2. The Spearman-Brown prophecy formula was used to estimate the KR_{20} of a forty item subtest from each of the subset KR_{20}s observed for the reading vocabulary subtest. In other words, elimination of the twenty-three items from the subtest did nothing to improve the average item reliability. A similar result was found for the reading comprehension reliability analysis. The KR_{20} coefficients are presented in Table 3.

Clearly, the results of the reliability analyses were counter to what we had expected. The elimination of items with either poor overall fits or poor fits for students achieving at a certain level should have resulted in item subsets with substantially higher average item reliabilities. That this was

Table 1

Retained Item Statistics

	Item	p-Value	Between t-Fit[a]	Total t-Fit[a]
Reading Vocabulary	1	.904	2.87	-1.41
	6	.823	1.97	-1.05
	9	.663	1.27	-2.45
	12	.599	2.57	0.18
	13	.548	1.44	2.20
	14	.771	1.47	0.83
	17	.592	2.42	-2.98
	20	.584	-0.04	1.88
	23	.459	1.30	1.60
	24	.602	1.91	-3.26
	25	.503	-0.62	-1.34
	26	.577	1.41	1.54
	27	.443	2.32	1.66
	31	.486	1.30	-2.73
	37	.417	2.30	0.56
	38	.384	2.05	1.47
	40	.495	1.81	-0.44
Reading Comprehension				
Passage # 1	1	.617	1.80	0.15
(8 items)	2	.754	1.81	1.06
	3	.552	0.08	1.83
	4	.670	-2.13	-0.33
	5	.729	2.29	-1.29
	6	.745	2.90	-1.11
Passage # 2	10	.610	1.25	-0.01
(7 items)	14	.606	1.11	-0.27
	15	.440	2.53	1.09
Passage # 3	16	.682	1.48	-2.97
(5 items)	18	.569	-1.65	0.46
Passage # 4	21	.601	-0.69	0.35
(7 items)	23	.518	-0.08	-1.54
	24	.525	2.62	-2.39
	25	.519	2.63	1.24
Passage # 6	32	.502	-1.22	1.05
(7 items)	33	.482	0.69	-1.21
Passage # 7	37	.583	2.77	-3.95
(7 items)	38	.541	2.85	-1.62
	39	.326	2.20	1.74
	40	.357	2.19	1.81
Passage # 8	42	.489	1.67	-2.36
(4 items)	44	.466	1.31	1.84

[a] t-fit values in this table are one-way statistics; only values significantly larger than zero indicate the presence of a poor fit.

Table 2

	All Students	Students Below 40th Percentile
Reading Vocabulary Subtest (40 items)	.924	.646
Reading Vocabulary Subset (17 items)	.836	.446

Table 3

	All Students	Students Below 40th Percentile
Reading Comprehension Subtest (45 items)	.944	.597
Reading Comprehension Subset (23 items)	.892	.443

not the case, we could only attribute to the fact that the criteria used to select the items was based on comparisons between an observed response matrix and a hypothetical response matrix, while the reliability analysis is based upon the inter-item correlations. Low achieving students will tend to have low p-values, which will in turn tend to suppress the values that the inter-item correlations can achieve. On the other hand, the item-fit statistics associated with the Rasch scaling procedure are not nearly as sensitive to low p-values.

Consequently, our second set of analyses to assess the item subsets were based on person-fit statistics from the BICAL analysis. Person fit, like item fit, is a t-statistic based on the disparity between a person's responses and the hypothetical probabilities generated for that person by the Rasch model.[4] In addition to computing an average person fit for the subtest and for the subset, average person fits were also computed within three groups: students who scored below the 40th percentile, students who scored between the 40th and 70th percentiles, and students who scored above the 70th percentile. The average t-statistics for person fit for reading vocabulary and reading comprehension are presented in Tables 4 and 5. Based on these analyses, the selection criteria would appear to have been appropriate. For both reading vocabulary and reading comprehension, the average fit statistic for the low achieving student has been substantially improved. For both subtests, the overall average person fit has also been improved, although for reading comprehension, the item subset had somewhat higher average person fits for students scoring above the 40th percentile. All of this would seem to indicate that both item subsets are capable of providing substantially better measurement of low achieving

---

[4]A person-fit t-statistic of zero indicates perfect fit.

Table 4

Reading Vocabulary

	All Students	Students Below 40th Percentile	Students Between 40th and 70th Percentiles	Students Above 70th Percentile
Reading Vocabulary Subtest	.35	.59	.22	.12
Reading Vocabulary Subset	.24	.32	.23	.11

Table 5

Reading Comprehension

	All Students	Students Below 40th Percentile	Students Between 40th and 70th Percentiles	Students Above 70th Percentile
Reading Comprehension Subtest	.25	.42	.16	.06
Reading Comprehension Subset	.22	.28	.22	.11

students, while continuing to provide accurate measurement of students in the middle and high achieving groups. Furthermore, because both subsets covered a substantial range of the item difficulties, the item subset scores can be equated to virtually the entire range of the subtest scores.

Conclusion

Based on these preliminary analyses, there appears to be some support for the concept of assessing student achievement with a subset of items that appear to be more appropriate for assessing the performance of low achieving students. Although the item subsets did not result in an improved average item reliability, the person-fit analysis clearly indicates that the item subsets are doing a better job of measuring low achieving students. We can now take these item subsets, and using the information from the Rasch scaling of the entire subtest, equate scores on the subset of items to scores on the entire subtest. Since the scores on the subtest are associated with norm-referenced scores, establishing this equating will permit anyone examining only the item subset scores to associate those scores with a norm-referenced interpretation. As a result, it will be possible to use only a subset of items to evaluate Title I programs, and concurrently have the results with respect to this subset of items referenced to performance by a nationally representative norming sample, as required by the Title I Evaluation and Reporting System. This means that Title I evaluators can use a statewide testing program for their evaluation and still be assured of having accurate information regarding the true achievement level of Title I students.

There is one other application of this Rasch scaling approach that bears mentioning. We can envision a number of situations where it might be useful

to examine the performance of a student or a group of students on only those items which are assessing a particular content area, such as the addition of single digit numbers. After a subset of items is selected using such a criterion, the Rasch scaling model could then be used first to assess how well the items seem to be measuring the same thing (such as achievement with respect to the addition of single digit numbers) and second to equate performance on this subset of items to performance on an entire subtest so that the item subset score can be given a norm-referenced interpretation. In the current study, we are creating a norm-referenced functional level test from a norm-referenced test; if item content is used as the selection criterion, then we are creating a norm-referenced criterion-referenced test from a norm-referenced test. The effect of this would be to enable the information gathered from a state- or district-mandated testing program to be used in a variety of ways, depending upon the information needs of the user. In this way, not only is the user supplied with relevant and appropriate information, but the data collection (i.e., test administration) burden is kept to an absolute minimum. Consequently, a variety of information needs can be met without having to raise the concern of excessive testing detracting from the delivery of services to students.

References

Forbes, D.W., & Ingebo, G.S. An empirical test of the content homogeneity assumption involved in Rasch item calibration. Paper presented at the meeting of the American Educational Research Association, Washington, D.C., 1975.

Specht, D.A. User's guide to subprogram RELIABILITY and repeated measurements analysis of variance. Ames, Iowa: Iowa State University, Department of Sociology, 1975.

Rasch. G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institut., Nielsen & Lydiche, 1960. (English translation)

Ryan, J.P., & Hamm, D.W. Practical procedures for increasing the reliability of classroom tests by using the Rasch model. Paper presented at the meeting of the National Council on Measurement in Educaiton, San Francisco, April 1976.

Wright, B.D., Mead, R.J., & Bell, S.R. BICAL: Calibrating items with the Rasch Model. Research Memorandum No. 23B. Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1979.

NTS RESEARCH CORPORATION

NTS Research Corporation is a private corporation specializing in research and development, policy analysis, test scoring services, and market and survey analysis for Federal, state and local agencies as well as business and industry. Founded in 1973, NTS is a rapidly growing and vital company that reflects the research and development community in which it exists. Located in Durham, North Carolina, near Duke University, the University of North Carolina at Chapel Hill, North Carolina State University, and the Research Triangle Park, a nationally renowned research center, NTS complements its own highly skilled professional staff with the services of both local and national experts.

NTS has over 90 staff organized into two major divisions, Social Science and Research Science, with seven operating centers, including Educational Research, Human Development, Health Studies, Operations Research and Data Processing, Statistical and Sampling Research, Market Research and Survey Operations, and Measurement and Data Management. Activities of these centers span the full range of contract research including program evaluation, technical assistance and training.

NTS corporate structure, management, personnel, and diverse resources contribute to the successful performance of all contracts. The strength and breadth of NTS corporate capability is represented by an experienced, highly qualified staff. In addition to advanced educational degrees, the staff possesses a range of technical and interpersonal skills as well as extensive experience in current social and educational issues. Through this staff, NTS combines rigorous scholarship with an active problem-solving orientation in its training, technical assistance, and research and development activities.

Below is a brief listing of some recent and continuing NTS contracts:

- Educational Policy Development Center for Educational Quality and Improvement.
  Office of the Assistant Secretary for Education/Policy Development.
- Evaluation of the Fund for the Improvement of Post-Secondary Education.
  Department of Health, Education and Welfare.
- Evaluation of State Dissemination Grants Program for Capacity Building.
  National Institute of Education.
- Evaluation of Urban Area Experience-Based Career Education Program.
  Office of Career Education, US Office of Education.
- Follow-Up Study of Patients Who Had I-131 and Other Diagnostic Procedures During Childhood.
  DHEW/FDA Bureau of Radiological Health.
- National Evaluation of Headstart Educational Services.
  Administration for Children, Youth and Families.
- Operation of the ESEA Title I Technical Assistance Center, HEW Region III.
  Office of Evaluation and Dissemination, US Office of Education.
- Delaware Item Pool Development.
  Delaware State Department of Public Instruction.
- Evaluation of the Comprehensive Project in Adult Functional Literacy Education.
  North Carolina State Board of Community Colleges.
- ESEA Title I Evaluation: District of Columbia.
  District of Columbia Public Schools.
- Educational Policy Development Center for Educational Quality and Improvement.
  Office of the Assistant Secretary for Education/Policy Development.